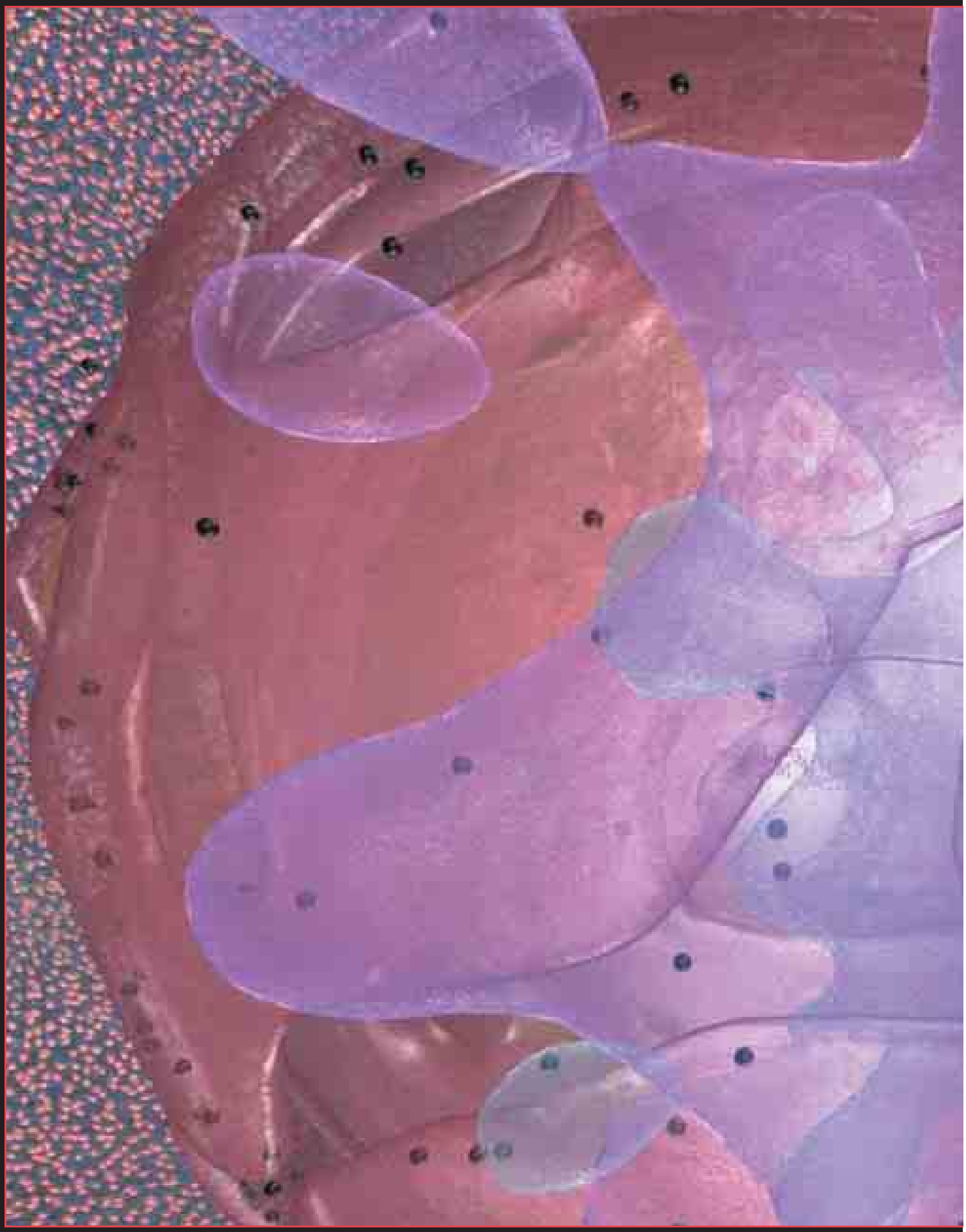

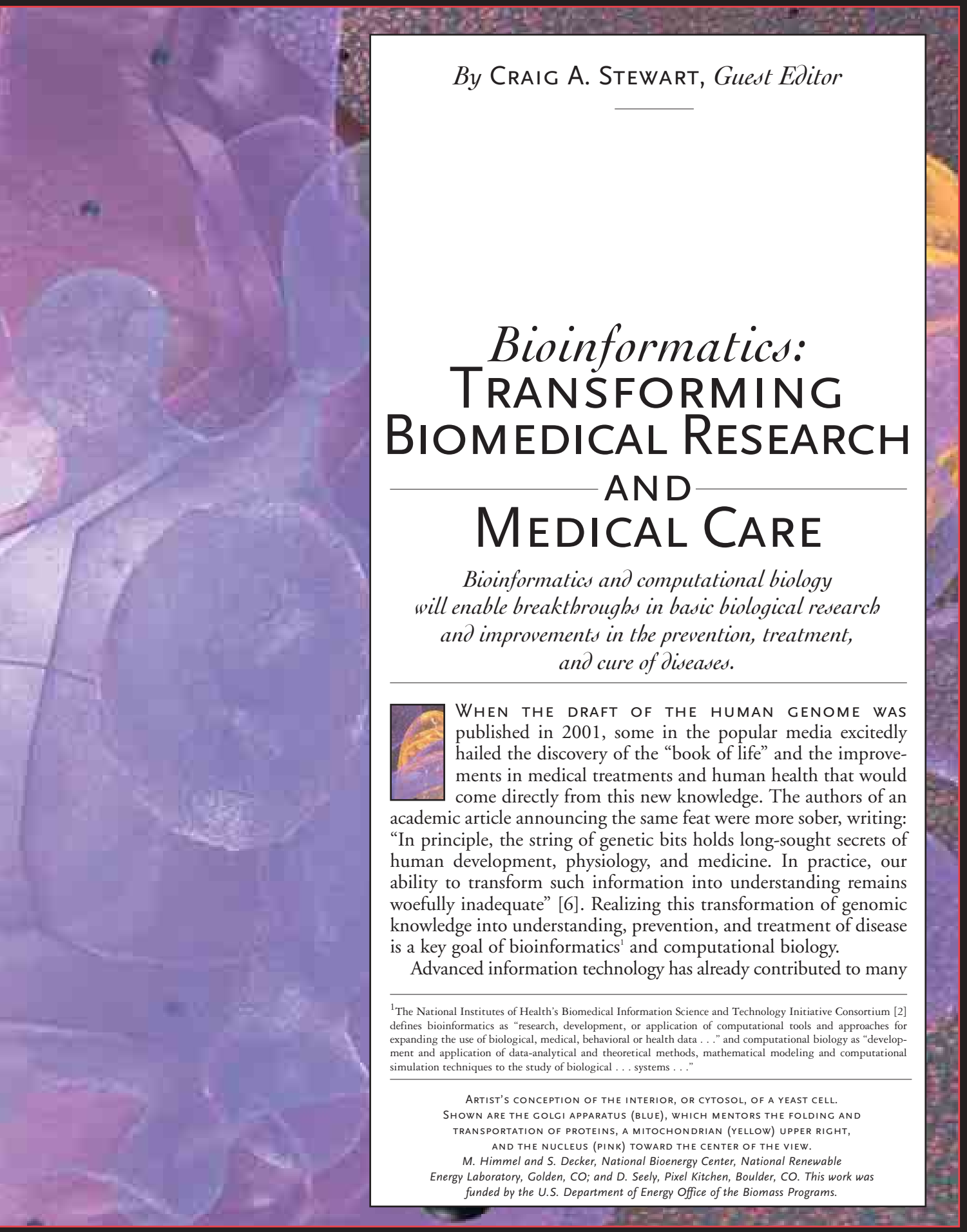

By Craig A. Stewart, *Guest Editor*

Bioinformatics: Transforming Biomedical Research and Medical Care

Bioinformatics and computational biology will enable breakthroughs in basic biological research and improvements in the prevention, treatment, and cure of diseases.

When the draft of the human genome was published in 2001, some in the popular media excitedly hailed the discovery of the "book of life" and the improvements in medical treatments and human health that would come directly from this new knowledge. The authors of an academic article announcing the same feat were more sober, writing: "In principle, the string of genetic bits holds long-sought secrets of human development, physiology, and medicine. In practice, our ability to transform such information into understanding remains woefully inadequate" [6]. Realizing this transformation of genomic knowledge into understanding, prevention, and treatment of disease is a key goal of bioinformatics[1] and computational biology.

Advanced information technology has already contributed to many

[1]The National Institutes of Health's Biomedical Information Science and Technology Initiative Consortium [2] defines bioinformatics as "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data . . ." and computational biology as "development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological . . . systems . . ."

Artist's conception of the interior, or cytosol, of a yeast cell. Shown are the golgi apparatus (blue), which mentors the folding and transportation of proteins, a mitochondrian (yellow) upper right, and the nucleus (pink) toward the center of the view. *M. Himmel and S. Decker, National Bioenergy Center, National Renewable Energy Laboratory, Golden, CO; and D. Seely, Pixel Kitchen, Boulder, CO. This work was funded by the U.S. Department of Energy Office of the Biomass Programs.*

SIMULATING THE ACTIVITY OF A SINGLE PROTEIN, TAKING INTO ACCOUNT EACH ATOM IN THE PROTEIN, *would take months using a PetaFLOPS-class computer* (ONE CAPABLE OF PERFORMING 10^{15} CALCULATIONS PER SECOND).

improvements in health care and to our understanding of basic biological processes. Supercomputers and advanced visualization techniques are used in preparation for brain surgery in ways that transform surgeons' ability to plan for high-risk operations. Supercomputers made possible the development of real-time functional MRI scans and are being used to improve radiation therapy in cancer treatment. Treatments for HIV infections can now be based on the specific genetic sequence of the particular virus infecting a patient. In the future, it will be routine to customize medical treatments to the genetic profiles of patients, their pathogens, or both. There is tremendous promise in this approach, as well as an economic consequence: as drugs become tailored to patients' genetic profiles, the market for each tailored drug shrinks. Such genetically tailored therapies will be economically viable only if drug development becomes faster and less costly. Faster and less-costly drug design can be accomplished through the use of high-performance computing. This approach will also make it possible for drug companies to more readily develop treatments for diseases that predominantly affect people living in economically disadvantaged areas or conditions.

Here, I outline the challenges that must be overcome before bioinformatics and computational biology are routinely employed to improve medical care and human health. Approaches to these challenges are detailed in the articles in this special section.

Biomedical science has been revolutionized by the development of robotic genetic sequencing systems and other analytic instruments that enable the creation of new biological data at an unprecedented rate. One of the fundamental challenges in bioinformatics is creating tools and techniques that permit the transformation of genetic sequence data into understanding of biological processes.

There is a natural fit between the problems of managing biomedical data and grid computing. Biomedical data sets are produced locally and used globally. Unique data sets are created in many individual laboratories, and a worldwide community of researchers wants to use these data sets as quickly after they are created as possible. A patient's medical record is updated and expanded in many locations but must be available in its entirety wherever and whenever the patient requires medical attention. New grid systems will create solutions to these needs. U.S. cyberinfrastructure initiatives strive to create grids of advanced storage, computation, and visualization systems linked to each other and to advanced instruments by high-speed networks (such as the TeraGrid, www.teragrid.org). Important efforts are under way in many other countries as well, including Canada, Japan, Taiwan, and the European Union (such as the U.K. e-Science Programme, www.rcuk.ac.uk/escience).

Grid computing is also well suited to the problem of making drug design faster, better, and less costly. Accomplishing this improvement requires the ability to accurately model how a new drug will affect the many patients who might take it. These models must take into account genetic variation among patients and enable accurate prediction of effectiveness and risks. This is a tremendously difficult modeling task because our bodies are incredibly complex, consisting of roughly 50 trillion cells. Each cell has on the order of 10^{13} components, many of which are proteins. Simulating the activity of a single protein, taking into account each atom in the protein, would take months using a PetaFLOPS-class computer (one capable of performing 10^{15} calculations per second). No such computer systems exist today, and designing one remains a tremendous challenge.

Given the massive computing power required to simulate the detailed behavior of a single molecule, it is clear that models of an entire organism cannot be based on every-atom or every-molecule approaches. Each of us is composed of a hierarchy of organ systems, organs, tissues, and cells. It should be possible to create a system of multiscale models that match this hierarchy, linked in ways analogous to the body's internal communications systems. These models are governed by the constraint that under normal circumstances our bodies function in ways that keep us alive. Implementing such models on a computing grid provides the means to employ computational resources more powerful than any supercomputer in existence today. It also offers the possibility of accurately predicting the activity of

potential new drug compounds before any laboratory testing is conducted.

The challenge in coming years will be to move from proofs of concept (such as improved surgical planning and better treatment for particular diseases) to a steady stream of broadly applicable successes in biomedical research and clinical practice. Such practical results will be possible only if there is a dramatic increase in the use of advanced computing tools in biomedical research and clinical medical practice, as called for in the National Institutes of Health Roadmap [8]. Biomedical applications will be a key driver of increased demand for high-performance computing and storage systems, advanced visualization environments, and high-speed networks. The computer science and biomedical communities can thus benefit from increased collaboration.

The articles in this special section describe current and important activities led by computer and biomedical scientists. The global nature of this endeavor is reflected in the diversity of authors, who represent six countries and three continents.

David A. Bader explains how several new grand challenges in life science (such as understanding protein folding and the process of evolution) require high-performance computing, focusing on the practical and theoretical challenges of implementing very large-scale computing systems. (Bader is a co-organizer of the IEEE International Workshop on High-Performance Computational Biology, www.hicomb.org, one of the key meetings for people using advanced computing techniques in biomedical research.)

Toshikazu Ebisuzaki et al. discuss two approaches to the creation of PetaFLOPS-scale computing systems. Building such systems is a significant challenge in computer engineering. But the result will enable new understanding of proteins, the most important building blocks of the cell.

Ross Overbeek et al. discuss the SEED, a new toolkit for organizing and understanding the vast and growing collection of publicly available genomic data, using peer-to-peer computing to accelerate fundamental biological research.

Mark Ellisman et al. explore four exemplars of biomedical computing grids now beginning to deliver practical benefits to biological researchers, clinical physicians, and, most important, patients. They range from computational grids for studying evolutionary history to data grids supporting better medical records management and improved diagnosis.

Homa Javahery et al. urge developers of bioinformatics tools to embrace the techniques of user-centered design to make computational biology tools more accessible to practicing biologists. Improved interface design will be a critical factor in encouraging and enabling greater use of bioinformatics tools and Web-based bioinformatics data sources.

Chris R. Johnson et al. discuss current best practices and future requirements for biomedical computing and visualization problem-solving environments. These environments permit multidisciplinary teams to work together to create tools based on the most accurate biological models possible and employ the most effective computer science applications available.

Improved human health is a great and pressing need. People who might otherwise die may instead lead normal, healthy lives, as the computer science and biological communities join forces to create new technologies for improving medical science and medical service delivery. Those who want to help make the promise of bioinformatics and computational biology a reality will find starting points in the articles here and in a number of books: [5] for bioinformatics; [1, 4] for grid computing; and [3, 7] for computational and systems biology. C

References

1. Berman, F., Fox, C., and Hey, A., Eds. *Grid Computing: Making the Global Infrastructure a Reality.* John Wiley & Sons, Ltd., West Sussex, England, 2003.
2. Biological Information Science and Technology Consortium Definition Committee. National Institutes of Health working definition of bioinformatics and computational biology (2000); www.bisti.nih.gov/compubiodef.pdf.
3. Bock, G. and Goode, J., Eds. 'In silico' simulation of biological processes. In *Novartis Foundation Symposium 247* (Novartis Foundation, London, Nov. 27–29). John Wiley & Sons, Ltd., West Sussex, England, 2002.
4. Foster, I. and Kesselman, C., Eds. *The Grid 2.* Elsevier, San Francisco, 2004.
5. Gibas, C. and Jambeck, P. *Developing Bioinformatics Computer Skills.* O'Reilly & Associates, Inc., Sebastopol, CA, 2001.
6. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature 409* (Feb. 15, 2001), 860–921.
7. Kitano, H., Ed. *Foundations of Systems Biology.* MIT Press, Cambridge, MA, 2001.
8. National Institutes of Health. *NIH Roadmap: Accelerating Medical Discovery to Improve Health.* Bethesda, MD, 2004; nihroadmap.nih.gov.

Craig Stewart (stewart@iu.edu) is the director of Research and Academic Computing, University Information Technology Services, and the director of the Indiana Genomics Initiative Information Technology Core, Indiana University.

Indiana University research in computer and life sciences is supported by: the National Science Foundation (Grant 0116050); National Institutes of Health (Grant 1U24AA014818-01); IBM Inc.; and the Lilly Endowment, Inc. (through its support for the Indiana Genomics Initiative of Indiana University). Any opinions, findings, and conclusions or recommendations expressed herein are those of the author and do not necessarily reflect the views of these organizations.